A cross-institutional, FAIR VIVO for Metabolomics

Michael Conlon, Kevin S. Hanson, Taeber Rapczak,
Naomi Braun, Christopher P. Barnes, University of Florida, Gainesville, Florida, USA

Metabolomics

Metabolomics is the scientific study of metabolites present within an organism, cell, or tissue. Human metabolites are small molecules found in human tissue that occur naturally as a result of human metabolism, or are present as a result of drugs, food components, or exposure to environmental conditions. Along with genomics (the study of DNA), transcriptomics (RNA), proteomics (proteins), metabolomics provides information regarding biochemical compounds and processes in cells, leading to a better understanding of cellular biology. “Metabolic profiling can give an instantaneous snapshot of the physiology of a cell, and thus, metabolomics provides a direct "functional readout of the physiological state" of an organism.” (Wikipedia http://bit.ly/2Pg06X3)

Such profiles can be used to distinguish tissue types, disease states, and health status of individuals. Understanding the changes in profiles over time can lead to improved understanding of diseases such as cancer and diabetes, leading to potential treatments.

Over 450,000 human metabolites have been identified to date. Metabolomics is considered an emerging field. Technical challenges in compound identification and data analysis are significant.

The NIH Common Fund Metabolomics consortium originated in 2013 to help address these and other challenges. Fourteen investigators across the US are supported to advance metabolomics. Five are engaged in compound identification, seven in software tools and data analysis. The group at UCSD supports the Metabolomics Workbench, a data repository. The group at the University of Florida provides overall coordination for the consortium.

Metabolomics Data

The Metabolomics Workbench (MWB) (https://metabolomicsworkbench.org) is the National Metabolomics Data Repository of the US National Institutes of Health (NIH). Anyone can deposit data to the MWB. As of August 28, 2019, the workbench provides data from 982 publicly available studies. Another 205 studies are currently embargoed and will be available subject to their embargo dates. MWB develops and uses RefMet (http://bit.ly/2PkxY5p), a nomenclature for representing metabolites found using mass spectroscopy (MS) and nuclear magnetic resonance (NMR) techniques. Investigators upload study data to the workbench and provide metadata regarding themselves and their work. The MWB provides an API that can be used to access metadata about studies and investigators. MWB metadata has been mapped to an ontology developed by the authors to represent it as RDF and load it to VIVO.

PubMed (http://pubmed.gov) is an open access index to literature in metabolomics. Its API can be used to find and retrieve publication data regarding metabolomics investigators. Many groups use PubMed data in VIVO. PubMed data has been mapped to the VIVO Ontology (http://vivoweb.org/ontology/core)

Data regarding software used in metabolomics studies has been difficult to find. An index to such software will be created and curated by a group at the University of Colorado Anschutz Medical Campus funded by the NIH. This data will be in the form of a spreadsheet. The authors will use the Software Ontology (SWO) (http://www.obofoundry.org/ontology/swo.html) to represent the data as RDF and load the data into VIVO.

FAIR Data Principles

The FAIR Data Principles (Wilkinson, et. al https://www.nature.com/articles/sdata201618) provide a framework for creating data that can be used by groups beyond the group that created the data. The principles are difficult to achieve in practice and much has been written about implementation. VIVO supports the principles naturally, it is designed to share data.

- Findable – data should be found using search engines on the Internet
- Accessible – the data can be accessed without technical, legal or operational barriers
- Interoperable – the data is available in a common format, so that data from more than one source can be readily combined
- Reusable – the data can be reused by those finding it. This typically means that enough information is provided for the recipients of the data to determine if thre data is suitable for their purposes.

The Metabolomics Workbench (https://metabolomicsworkbench.org), Metabolites (https://www.ebi.ac.uk/metabolights/index), and the MassBank (http://massbank.jp) provide data on experimental results, and spectra of compounds according to the FAIR principles. In each case, data can be found by using internal search capabilities of each of the sites. This is not ideal. Generally findable results using search engines can be created for metabolomic data sets using the techniques described here. Each dataset or study will have its own page which can be found and indexed by search engines. Our approach will result in pages for datasets, publications, investigators, and software.

Figure 1. Metabolomics.info web site shares information about events, and the NIH consortium

Figure 2. The Metabolomics Workbench is a data repository for data from metabolomics experiments. An API provides metadata regarding studies and investigators for all to use.

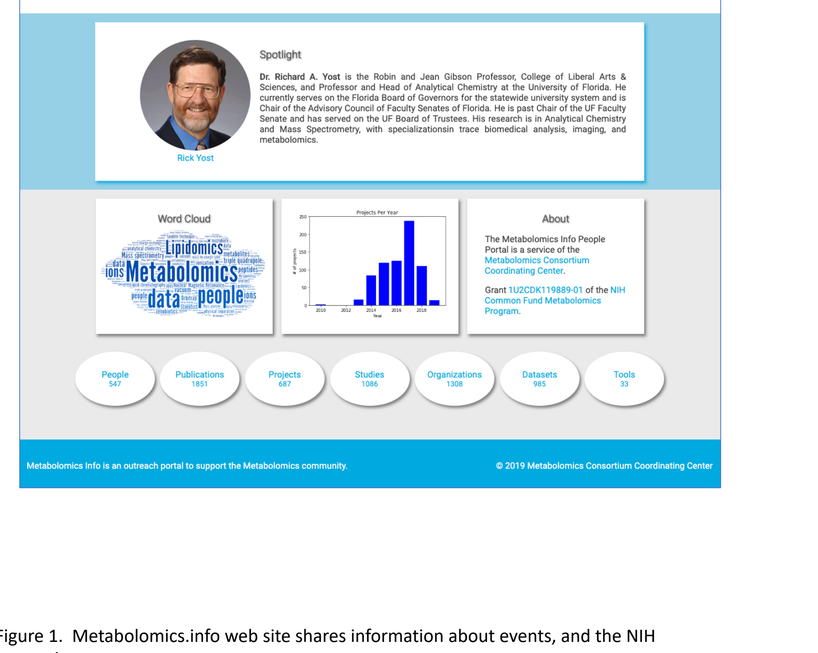


Figure 1. Metabolomics.info web site shares information about events, and the NIH consortium

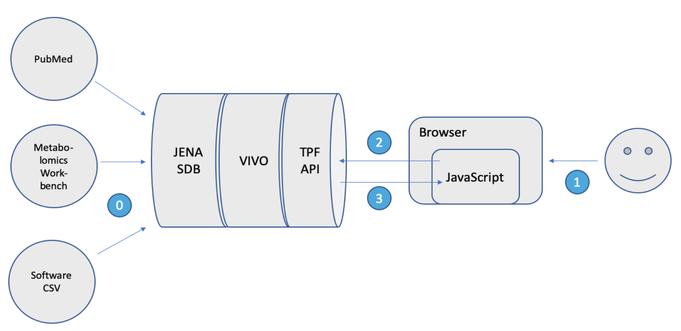
Use Cases

The metabolomics community is still developing. Some use cases for the web site being developed include:

1. Discover metabolomics datasets using search engines. The pages created by the approach described here.
2. Organize metabolomics information by investigator. Creating web pages for each investigator will emphasize the number of datasets deposited by each investigator, the number of metabolomics papers of the investigator, and software systems created by the investigator.
3. Provide summaries of contributions by investigator.
4. Provide contact information for investigators.
5. Boost the visibility of the work products (papers, software, and datasets) of the NIH Common Fund Metabolomics Consortium, leading to increased collaboration between consortium members and others around the world.
6. Increase deposits to Metabolomics workbench
7. Increase reuse of data in Metabolomics workbench
8. Through the FAIR data principles, encourage data sharing and reuse in the metabolomics community.

Technical Approach

See figure below. 0) Publication data from PubMed, dataset metadata from Metabolomics workbench, and software metadata from collaborators at the University of Colorado Anschutz Medical Campus are retrieved, transformed to triples and loaded to the JENA SDB triple store (part of VIVO) using custom scripts on a regular basis; 1) User makes request for page that will contain data from VIVO; 2) TPF Client JavaScript software makes a TPF query of the VIVO TPF endpoint. The JavaScript software is generic – any ontology and any TPF endpoint can be used; 3) the TPF API returns a set of triples to the client. No modifications of VIVO are necessary.



Ontology

A small ontology was created to represent the data for the metabolomics consortium application. The ontology simplifies the VIVO ontology and most assertions can be inferred from the VIVO ontology.

For example, in the VIVO ontology, is a person p is the pi of a grant g, the VIVO ontology would assert:

```
p bearerOf r
r a PIrole
r relatedBy g
```

A new object property `isPIOf` (“is principal investigator of”) is used to simplify the above assertions to the equivalent assertion below:

```
p isPIOf g
```

A TPF query of the form

```
p isPIOf *
```

returns all the grants for which p is the principal investigator. Other “shortcut” object properties include `isAuthorOf` for publications and `isCreatorOf` for datasets. Such shortcuts speed up the retrieval of metadata using TPF. The ontology is available here: https://github.com/ctsit/metabolomics_ontology

Triple Pattern Fragments (TPF)

Triple Pattern Fragments (TPF), (Verborgh, et. Al doi:10.1016/j.websem.2016.03.003) is a low cost knowledge interface for the web. TPF provides a simple, fast, general mechanism for querying knowledge graphs of RDF triples. TPF answers queries of the form:

Subject predicate object

Where any of the three components can be a wildcard. A query such as

```
* rdf:type *
```

returns all the type assertions in a triple store. When executed at http://openvivo.org/tpf/core, this query returns 3.8M triples. The query

```
<uri> * *
```

returns all triples with the specified uri as a subject. When executed on a person’s URI in VIVO, the resulting set of triples is the first order graph of the person’s information in VIVO.

A TPF endpoint is included as part of the VIVO installation in version 1.10 and above based on the the TPF server available here: https://github.com/LinkedDataFragments

JavaScript

A small JavaScript library is available (http://github.com/ctsit/tpf) to make TPF queries and handle TPF replies. The code below adds the library to a page, and creates a client pointing at the OpenVIVO (http://openvivo.org) TPF endpoint. The client then requests the triples for a particular entity. The triples returned by the query are then used in a second query (`.Link`) to return all the `rdfs:label` values from the first query. The `.Single` function provides a callback function that executes on one returned value when the queries are completed. In the example, the callback displays a single label in the browser console log.

```
<script href="tpf.js"></script>
<script>
    const rdfs = "http://www.w3.org/2000/01/rdf-schema#"
    const endpoint = "http://openvivo.org/tpf/core"
    const client = new tpf.Client(endpoint)

    client
        .Entity("http://openvivo.org/a/orcid0000-0002-1304-8447")
        .Link(rdfs, "label")
        .Single(function (label) { console.log(label) })
</script>
```

Additional functions in the library include: 1) `.Results` provides a callback that can process a list of returned results. 2) `.Query` can be used to specify an arbitrary subject, predicate, object TPF query, returning a set of triples. 3) `.Type` is used to reduce a returned set of triples to entities of a specified `rdf:type`.

Progress and Next Steps

TPF is being used to develop web pages for metadata regarding metabolomics coming from multiple sources – Metabolomics Workbench, PubMed, and a spreadsheet describing metabolomics software tools.

The approach described here is easily extended to support:

1. **Internationalization.** The client can accept a language parameter that is used to filter triples to those containing the preferred language, and any secondary languages desired.
2. **Cross-site TPF.** A web page can have any number of clients, each client associated with a specified TPF endpoint. References in one VIVO to triples in a second VIVO can result in the the creation of a second client for the second VIVO and retrieval of desired triples from the second VIVO using TPF.

This poster is the work of the Metabolomics Consortium Coordinating Center (M3C) at the University of Florida and is funded in part by the grant 1U2CDK119889-01 of the NIH Common Fund Metabolomics Program.
The content of this poster is solely the responsibility of the M3C and does not necessarily represent the official views of the NIH.

UF | UNIVERSITY of FLORIDA